Model selection (3)

Manel Martínez-Ramón

ECE, UNM

October 2018

Criteria for parameter adjustment

- The most common parameter estimation method consists of a validation procedure. But this may become unpractical in two also common situations:
  - When the data set is small.
  - When the number of hyperparameteres is high.
- Therefore, we must explore other selection criteria that have their roots in a probabilistic perspective.

Marginal likelihood

- We can use the expression of the marginal likelihood of $\mathbf{y}$ as a function of the estimator $\mathbf{f}$, where we ignore the possibility of using different structures:

  $$p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = \int_{\mathbf{f}} p(\mathbf{y}|\mathbf{X},\mathbf{f},\boldsymbol{\theta})p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})d\mathbf{f}$$

  Here we assume that $\mathbf{f}$ is constructed with functions with parameters $\boldsymbol{\theta}$.

Marginal likelihood

- In particular, $\mathbf{f}$ is a Gaussian process whose covariance matrix has been constructed with kernel functions with these parameters.

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{K})$$

Also, the likelihood of $\mathbf{y}$ is a Gaussian with mean $\mathbf{f}$ and covariance $\sigma_n^2\mathbf{I}$

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{f}, \sigma_n^2\mathbf{I})$$

Marginal likelihood

- If we put both distributions in the integral, we find that the result is another Gaussian distribution. We are only interested in the exponent of the Gaussian, so we compute the logarithm of this function:

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\top \mathbf{K}_\mathbf{y}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}_\mathbf{y}| - \frac{n}{2}\log 2\pi$$

where

$$\mathbf{K}_\mathbf{y} = \mathbf{K}_\mathbf{f} + \sigma_n^2\mathbf{I}$$

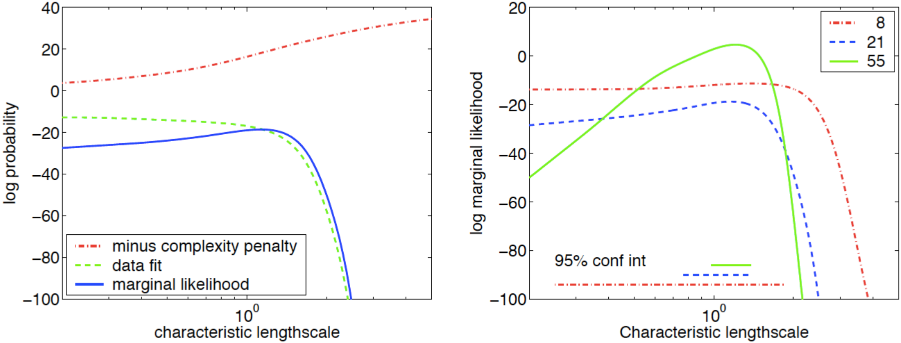
and $\mathbf{K}_\mathbf{f} = \mathbf{K}_\mathbf{f}(\boldsymbol{\theta})$

Marginal likelihood

The terms of the log-likelihood are fully interpretable:

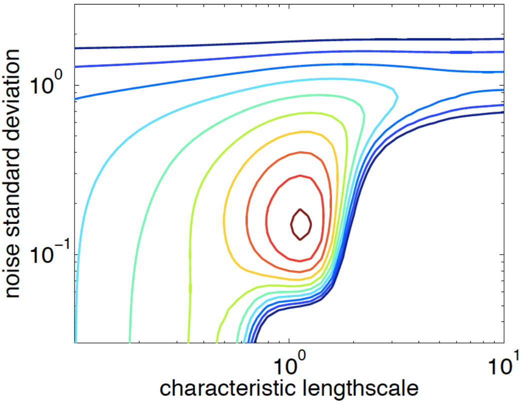
- Term
  $$-\frac{1}{2}\mathbf{y}^\top \mathbf{K}_\mathbf{y}^{-1}\mathbf{y}$$
  depend on the regressors. It is the equivalent to the empirical error.
- Term
  $$-\frac{1}{2}\log|\mathbf{K}_\mathbf{y}|$$
  does not depend on the regressors and it represents the complexity used in the input data representation.
- $-\frac{n}{2}\log 2\pi$ is only a normalization term.

Marginal likelihood



Left: Empirical term and complexity as a function of the SE kernel width of example in video 8.2. Right: Log marginal likelihood as a function of the SE width for different sizes of training sets. Taken from Rasmussen et al. 2006.

Marginal likelihood



Log likelihood as a function of the width and the noise parameter. Taken from Rasmussen et al. 2006.

Gradient of the marginal likelihood

- Note that there is a clear maximum of the likelihood in the point where the optimal parameters are.
- Clearly, we can compute the gradient of the log-likelihood as a function of the hyperparamenters in order to find the maximum. For a given parameter θ_j we find

$$\frac{\partial}{\partial \theta_j} \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \frac{1}{2} \mathrm{tr}\left(\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - \mathbf{K}^{-1} \right) \frac{\partial \mathbf{K}}{\partial \theta_j} \right)$$

where $\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y}$

Gradient of the marginal likelihood

- The optimization can be performed using a gradient ascent procedure:
$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \mu \Delta_{\boldsymbol{\theta}} \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$$
- Of course, this is not guaranteed to have a single minimum, so different initialitations have to be applied.
- Also, we perform this optimization over the training data. This is somewhat prone to overfitting. Thus, a Leave One Out cross validation procedure is recommended.

LOO

- A LOO is performed by computing the log-likelihood of a sample taken out of the training.
- So for each sample y_i we construct a data set $\mathbf{X}_{-i}, \mathbf{y}_{-i}$ where we take out sample $\mathbf{x}_i, y_i$ to be used as validation set.
- Then we compute the likelihood of this sample, and its derivative, and we average for all samples. The log-likelihood is

$$\log p(y_i|\mathbf{X}_{-i}, \mathbf{y}_{-i}, \boldsymbol{\theta}) = -\frac{1}{2}\log\sigma_i^2 - \frac{y_i - \mu_i^2}{2\sigma_i^2} - \frac{1}{2}\log 2\pi$$

That is obtained using the predictive likelihood of a test data.

LOO

- In expression

$$\log p(y_i|\mathbf{X}_{-i}, \mathbf{y}_{-i}, \boldsymbol{\theta}) = -\frac{1}{2}\log \sigma_i^2 - \frac{y_i - \mu_i^2}{2\sigma_i^2} - \frac{1}{2}\log 2\pi$$

we must compute the mean μ_i and the variance σ_i^2 from the equations of the predictive posterior, that can be reduced to

$$\mu_i = y_i - [\mathbf{K}^{-1}\mathbf{y}]_i / [\mathbf{K}^{-1}]_{ii}$$

and

$$\sigma_i^2 = 1/[\mathbf{K}^{-1}]_{ii}$$

LOO

- After tedious algebra, we can find the expression of the sum of the derivatives of the predictive log-likelihood wrt the LOO samples

$$\frac{\partial L_{LOO}}{\partial \theta_j} = \sum_{i=1}^{N} \left(\alpha_i [\mathbf{Z}_j \boldsymbol{\alpha}]_i - \frac{1}{2} \left(1 + | \frac{\alpha_i^2}{[\mathbf{K}^{-1}]_{ii}} \right) [\mathbf{Z}_j \mathbf{K}_{ii}^{-1}] \right) / [\mathbf{K}^{-1}]_{ii}$$

where $\mathbf{Z}_j = \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \theta_j}$. We can proceed in a similar way as in slide 10 to optimize the hyperparameters.

Outcomes of this lesson

In this lesson we have seen:

- The criterion for hiperparameter optimization, based on maximum likelihood.
- The procedure for the optimization, based on gradient ascent.
- The modification of this process, that includes a Leave-One-Out strategy to reduce the chance of overfitting.